# Explore In-Context Learning for 3D Point Cloud Understanding - Supplementary Material Paper ID: 4623

**Anonymous Author(s)**
Affiliation
Address
email

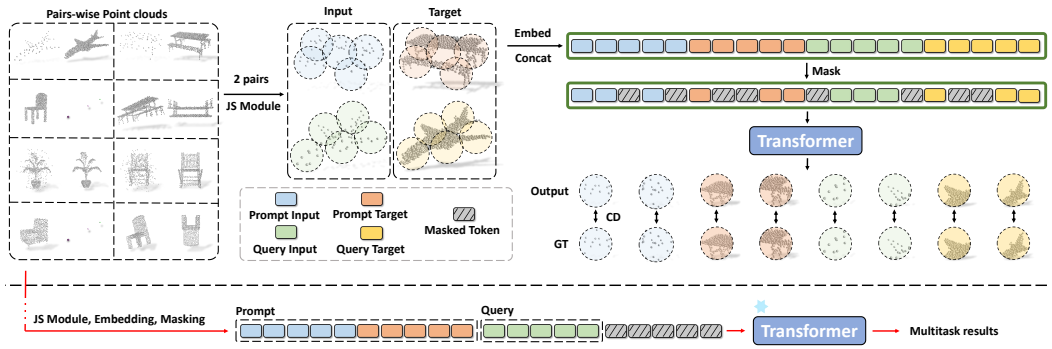


Figure 1: **Overall scheme of our Point-In-Context-Cat**. ***Top***: During training, each sample comprises two pairs of input and target point clouds that tackle the same task. Unlike PIC-Sep, PIC-Cat concatenates the input and target to form a new point cloud. ***Bottom***: In-context inference on multitask. Our Point-In-Context could infer results on various downstream point cloud tasks.

**Overview.** The supplementary material includes sections as follows:

- Section A: Pipeline of Point-In-Context-Cat, including training and inference stages.
- Section B: Additional ablation studies on PIC, including the prompt selection solutions, mask ratio, and the loss function.
- Section C: More results about multitask models trained using a pre-trained backbone with multitask heads.
- Section D: More visual results and corresponding analysis.

## A More Details of PIC

**Pipeline of PIC-Cat.** During the training phase, our approach involves selecting a pair of query point clouds and a pair of prompt point clouds from the training dataset. These point clouds are then grouped using the Joint Sampling module. Following this, we perform encoding and tokenization on each point cloud and concatenate them to create a new point cloud. A masking operation is applied to the entire point cloud to conduct the MPM task. We set the mask ratio as 60% for our specific approach, PIC-Cat. During the in-context inference stage, we only mask the last quarter of the tokens, which corresponds to the desired output. This approach allows our PIC-Cat model to reconstruct the masked tokens, leveraging its training experience. It is important to note that the task on the query point cloud is determined by the prompt in the example pair.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Table 1: The comparison of parameters, GFLOPs, and test speed.

| | Task-specific models | | | multi-task models | | | | In-context learning models | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PointNet [7] | DGCNN [8] | PCT [4] | PointNet [7] | DGCNN [8] | PCT [4] | Point-MAE [6] | Point-BERT [9] | PIC-Cat | PIC-Sep |
| Params(M) | 8.9 | 7.9 | 13.0 | 6.0 | 7.6 | 10.8 | 27.0 | 52.6 | 29.0 | 28.9 |
| FLOPs(G) | 1.9 | 3.1 | 6.3 | 1.9 | 10.2 | 6.4 | 11.8 | 12.0 | 12.1 | 8.4 |
| Test speed | 694 | 1500 | 694 | 844 | 1185 | 717 | 742 | 190 | 953 | 291 |

Table 2: More ablation study on Point-In-Context.

(a) Prompt selection method.

| Model | Selection method | Rec. CD↓ | Den. CD↓ | Reg. CD↓ |
|---|---|---|---|---|
| PIC-Cat | Fea-aware | 4.30 | 5.32 | 10.68 |
| PIC-Cat | CD-aware | **4.28** | **5.26** | **9.65** |
| PIC-Sep | Fea-aware | 4.90 | 7.57 | 5.79 |
| PIC-Sep | CD-aware | **4.38** | **7.06** | **4.12** |

(b) Ablation study on mask ratio.

| # | Mask Ratio | Rec. CD↓ | Den. CD↓ | Reg. CD↓ | Part Seg. mIOU↑ |
|---|---|---|---|---|---|
| 1 | 0.2 | 27.8 | 33.5 | 68.2 | 43.84 |
| 2 | 0.3 | 5.2 | 7.3 | 14.8 | 56.72 |
| 3 | 0.4 | 5.0 | 7.4 | 12.3 | 60.25 |

(c) Different loss functions.

| # | Loss function | Rec. CD↓ | Den. CD↓ | Reg. CD↓ | Part Seg. mIOU↑ |
|---|---|---|---|---|---|
| 1 | $\ell_1$ | 5.0 | 8.1 | 11.1 | 72.35 |
| 2 | $\ell_2$ | **4.7** | **7.6** | **10.3** | **74.95** |
| 3 | $\ell_1 + \ell_2$ | 5.3 | 7.9 | 13.3 | 70.46 |

**Comparison of Model Parameters, GFLOPs, and test speed.** We compared the parameters and GFLOPs of each model in the main results of the main text in Tab. 1. Our model achieves a favorable balance between structural complexity and task performance, making it a compelling choice. Note that the parameters and GFLOPs of task-specific models are computed, including four individual models for four different tasks. Besides, we report the speed of models by samples/second tested on one NVIDIA RTX 3080 Ti GPU. Our PIC-Cat presents a high inference speed (953 samples/second), which is second only to DGCNN [8].

# B More Ablation Studies

Given the length of the main text, we include additional ablation experimental results in this section. It is important to note that these experimental results hold equal significance to the main text.

**Prompt Selection.** The selection of the prompt significantly influences the quality of the PIC's output. Therefore, in addition to the random and class-aware prompts discussed in the main text, we further delve into selecting two alternative prompts. To select pairs of examples that are paired with the query point cloud, we consider two factors: the Chamfer Distance (CD) between the prompt and the query point cloud and the feature similarity between them (features are obtained from pre-trained PointNet [7]), which are respectively denoted as CD-aware and Fea-aware. By considering these criteria, we can identify the most suitable prompts to accompany the query point cloud, thus enabling PIC to achieve optimal performance.

As depicted in Tab. 2(a), the CD-aware method demonstrates the best performance than Fea-aware, surpassing even the individual models trained separately on the registration task. This finding highlights the effectiveness of the prompt selection approach in enhancing performance compared to the results presented in the main text. Similar findings can also be found in the 2D in-context learning framework [1, 10].

**Mask Ratio.** In addition to the high mask ratio discussed in the main text, we also conducted experiments on PIC-Sep with a low mask ratio ranging from 20% to 40%. As shown in Tab. 2(b), training PIC with a lower mask ratio weakens its performance across various tasks, especially on the mask ratio 20%. Different from language data, we also find keeping sparsity in training is necessary for mask point modeling for in-context learning. We find similar results as in MAE [5], a higher mask ratio is required to make sure that the model can learn hidden features well.

Table 3: Results of multitask models composed of a multitask head and a pre-train backbone trained on ShapeNet [3] for classification. For reconstruction, denoising, and registration, we report Chamfer Distance $\ell_2$ loss (x1000). For part segmentation, we report mIOU.

| Models | Acc.(%) | Reconstruction CD ↓ | | | | | | Denoising CD ↓ | | | | | | Registration CD ↓ | | | | | | Part Seg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | L1 | L2 | L3 | L4 | L5 | Avg. | L1 | L2 | L3 | L4 | L5 | Avg. | L1 | L2 | L3 | L4 | l5 | Avg. | mIOU↑ |
| multitask models: share backbone + multi-task heads | | | | | | | | | | | | | | | | | | | | |
| PointNet [7] | 88.7 | 47.0 | 45.8 | 45.4 | 45.4 | 45.8 | 45.9 | 22.9 | 23.2 | 26.3 | 28.3 | 30.0 | 26.1 | 35.5 | 34.8 | 37.1 | 37.2 | 38.6 | 36.6 | 10.13 |
| DGCNN [8] | 89.4 | 46.7 | 47.2 | 48.1 | 48.6 | 48.5 | 47.8 | 8.2 | 8.3 | 8.4 | 8.8 | 9.2 | 8.6 | 14.2 | 15.8 | 18.2 | 21.8 | 23.5 | 18.7 | 21.35 |
| PCT [4] | 89.5 | 64.7 | 60.8 | 59.2 | 60.1 | 59.7 | 61.0 | 14.5 | 12.2 | 12.4 | 12.0 | 11.8 | 12.6 | 22.6 | 25.2 | 28.3 | 31.1 | 33.2 | 28.1 | 15.43 |

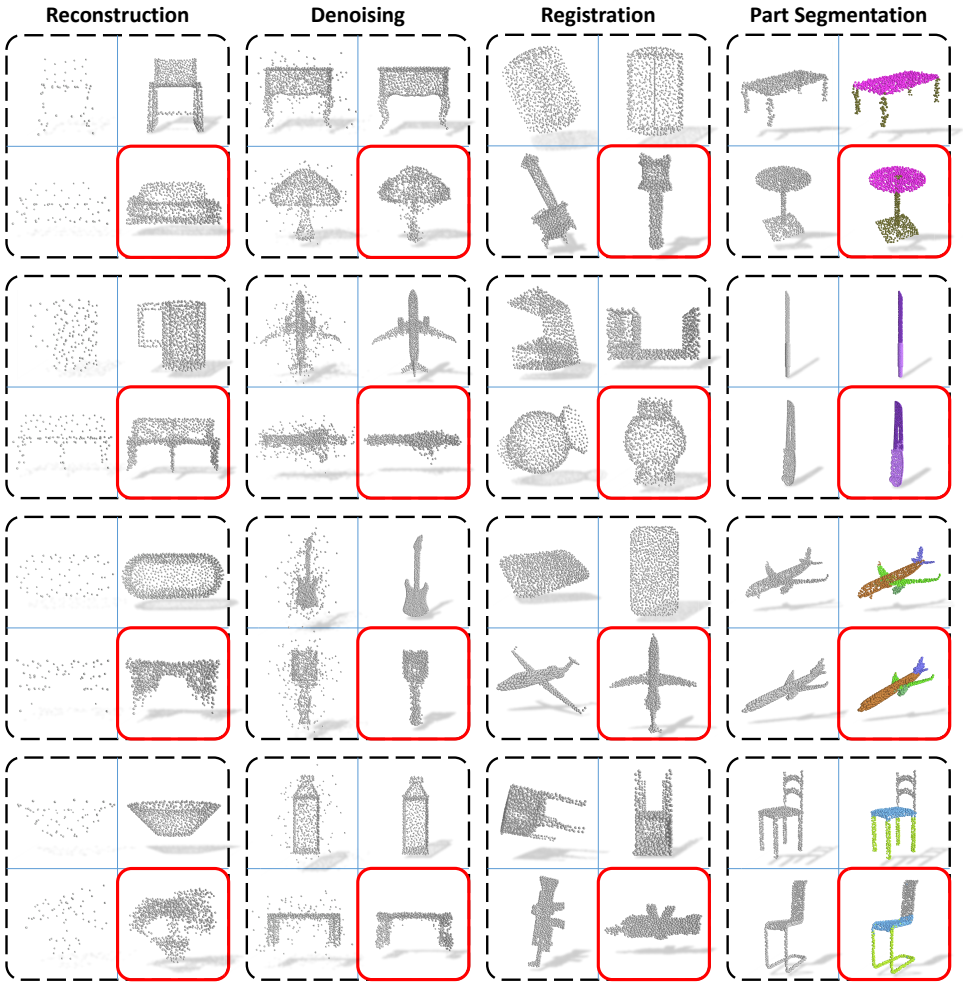


Figure 2: **Additional visualization results of PIC-Sep.** The output of our model is marked in red. Note that the results of part segmentation have been processed by adding XYZ coordinates.

**Loss Function.** We conducted an exploration to determine which loss function is most suitable for our PIC (Point-In-Context) model. During training, we experimented with using $\ell_1$, $\ell_2$, and a combination of $\ell_1$ and $\ell_2$ as the loss functions for our PIC-Sep. As Tab. 2(c) shows, $\ell_2$ achieves the best result on various tasks.

## C More Results of Multi-task Models

**Pre-trained Backbone + multitask Heads.** For multitask models, we utilize a pre-trained backbone feature extraction network that is trained on the ShapeNet [3] dataset for classification tasks. This pre-trained backbone network is equipped with multiple task-specific heads to perform multitask

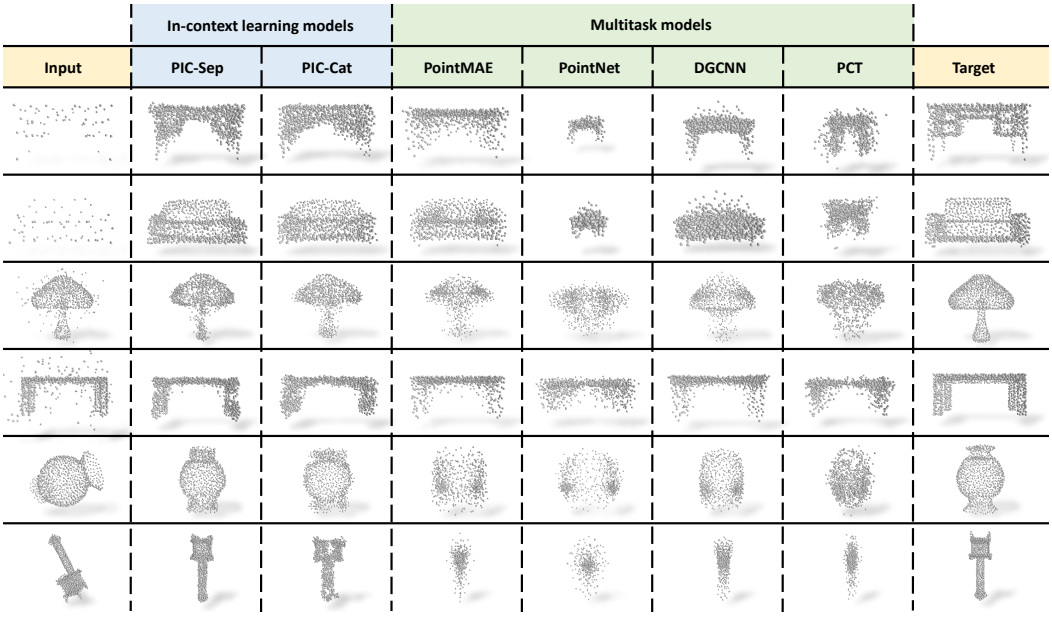


Figure 3: **Visualization of comparison results between PIC and multitask models.**

learning on our benchmark, allowing for the simultaneous handling of various tasks, including reconstruction, denoising, registration, and part segmentation. As shown in Tab. 3, while these supervised models perform well when trained on individual tasks, they exhibit poor performance on multitask benchmarks. Despite their success in isolated tasks, the models struggle to effectively handle multiple tasks simultaneously, resulting in subpar results in the context of multitask benchmarks.

# D More Visualization

**More visualization of PIC-Sep.** We visualize more examples in Fig. 2, including reconstruction, denoising, registration, and part segmentation.

**Comparison results between PIC and multitask models.** We conducted a comparison of visualization results between PIC (Point-In-Context) and multitask models on three tasks, including reconstruction, denoising, and registration. It is important to note that the multitask models in this comparison do not utilize the pre-trained backbone. As shown in Fig. 3, compared with other multitask models, our PIC-Sep and PIC-Cat output results are more satisfactory.

**Board impact.** Our work is the first to explore in-context learning in 3D point clouds, including task definition, benchmark, and baseline models. Due to the limited computation resource, we do not perform more experiments such as outdoor scene segmentation and large-scale point cloud datasets [2]. This will be our further work.

# References

[1] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *NeurIPS*, 2022.

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv:1512.03012*, 2015.

[4] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *CVM*, 2021.

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[6] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *ECCV*, 2022.

[7] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017.

[8] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *TOG*, 2019.

[9] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *CVPR*, 2022.

[10] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *arXiv:2301.13670*, 2023.